Legal Memorandum Generation Using Retrieval-Augmented Large Language Models and Dutch Case Law

MIHAI TIMOFICIUC, University of Twente, The Netherlands

Legal memo drafting is a critical but labor-intensive task in Dutch legal practice, requiring the synthesis of case law into structured, actionable documents. This thesis presents a Retrieval-Augmented Generation (RAG) system that generates legally grounded memoranda based on Dutch administrative court rulings and combines a curated corpus of 200 social security judgments with semantic chunking, pgvector-based retrieval, and GPT-4.1-driven memo generation. To enforce legal traceability, a structured intake form, and strict citation requirements were employed, combined with a modular evaluation framework developed to assess factual accuracy, citation precision, and semantic grounding using both automated heuristics and GPT-4.1 judgment. Results show the system achieves perfect citation precision (1.0), consistent recall (0.78), and F1 score (0.87) across multiple similarity thresholds. Reviewer LLMs added limited improvement, reinforcing the conclusion that robust retrieval and prompt design are more impactful than complex verification layers. To promote reproducibility and research advancement, the full implementation has been made available as open-source software [27].

Additional Key Words and Phrases: Retrieval-Augmented Generation, Legal AI, Dutch Case Law, Citation Verification, Hallucination Control, LLM

1 INTRODUCTION

Legal memos are structured written documents that succinctly identify a legal question, provide a thorough review of relevant statutes, regulations, and case law, and deliver actionable advice for clients, colleagues, or courts. The term "memo" is employed throughout this thesis as a concise reference to legal memorandum, reflecting both common legal practice and the accessible nature of automated document generation systems that prioritize brevity and clarity over formal terminology.

Legal professionals in the Netherlands use a vast and growing body of Dutch case law in their day-to-day work in order to inform their analysis. They manually locate the most relevant cases, parse lengthy judgments, and synthesize the reasoning into a coherent memo. This process is both time-consuming and error-prone, diverting hours of work from higher-value tasks such as client counseling.

Advancements in Retrieval-Augmented Generation (RAG) showcase a promising way of streamlining this process. By combining a semantic retrieval component, capable of pinpointing pertinent passages from a corpus of case law, with a Large Language Model, RAG systems can produce draft memos that put together facts, legal reasoning, and citations in a single workflow.

RAG applications in other knowledge-intensive domains (e.g., open-domain question answering) have demonstrated significant gains in factual grounding and citation fidelity compared to generation-only approaches. However, even with the help of RAG, Large Language Models (LLMs) still suffer from hallucinations, fabricating or misattributing legal rules or case references and producing inaccurate citations. These errors undermine trust in the generated output. This is detrimental for high-stakes legal settings, where even a minor factual error can have serious ethical and professional consequences.

In order for the RAG system to be useful, it must not only retrieve and assemble relevant case fragments, but also ensure that every assertion in the memo is traceable to a court ruling from the knowledge base. This paper proposes developing a proof-of-concept RAG system for generating legal memos based on a curated subset of Dutch case law extracted from `Rechtspraak.nl` (the website of courts, courts of appeal, and special colleges in the Netherlands). The corpus will focus on a collection of 200 judgments in a single legal domain, and the project will make use of both prompt engineering and post-generation verification steps to provide a balance between speed and reliability.

2 PROBLEM STATEMENT

The creation of legal memos in Dutch legal practices currently depends on case law retrieval and synthesis, which is a time-consuming process prone to oversight. Even though RAG architectures offer the potential to automate memo creation, LLMs produce hallucinations at a frequent rate. Such hallucinations materialize in fabricated or misattributed legal references and inaccurate citations. These factors erode user confidence and prevent system usability. Moreover, existing AI tools have not been specifically tailored to the Dutch jurisdiction or directly addressed the dual challenges of minimizing hallucinations and ensuring citation verifiability in legal memo creation. Without a targeted solution, adopting a practical AI-assisted memo generator in the context of Dutch legal practices remains constrained. Social security law was selected as the focus domain due to its high frequency in administrative court cases, relatively consistent reasoning structure, and clear applicability to everyday legal practice. This made the domain ideal for evaluating citation accuracy and hallucination control in automated memo generation.

Therefore the central research question is:

Can a RAG system be designed to produce legal memos with high factual accuracy, verifiable citations, and minimal hallucinations using a curated corpus of Dutch social security law rulings?

This research specifically focuses on the task of retrieving and generating legally grounded memos from rulings concerning social benefits disputes, evaluating factual accuracy and citation reliability through automated and optional human assessments.

3 RELATED WORK

The legal Natural Language Processing (NLP) field contains numerous studies that explore domain-specific language models, hallucination reduction, and Retrieval-Augmented Generation (RAG). Models such as Legal-BERT [2] and its Dutch variant RechtBERT [16] demonstrate that pre-training on legal corpora improves performance on legal tasks. However, the main challenges remain grounding and input length limits (see Appendix A.2 for full literature review notes). RAG has proven effective in increasing factuality and

TScIT 43, July 4, 2025, Enschede, The Netherlands

citation quality in high-stakes domains like law, with the combination of retrieval with generation improving accuracy compared to generation-only models in studies such as Lewis et al. [15], Redelaar et al. [24], and Pipitone et al. [22] with recent benchmarks such as LegalBench-RAG emphasizing snippet-level retrieval to reduce hallucination risk. Furthermore, the BART model for legal summarization [25] showed promise but suffered from omissions of legal context and citations with domain transferability [1]. Hallucinations persist despite RAG's improvements, especially when retrievals have low relevance [10, 21]. Recent frameworks such as *Self-Refine*[17] and *Chain-of-Verification*[7] prompt LLMs to revise their outputs, showing gains in factual accuracy. However, their specific effectiveness in the legal domain remains uncertain.

This thesis addresses a key research gap: while prior work has advanced legal-domain LLMs, RAG techniques, and methods for mitigating hallucination, no study has systematically evaluated the generation of Dutch legal memos using a structured RAG pipeline with grounded citations.

4 METHODOLOGY

This project follows a four-phase methodology combining literature review, legal data processing, system prototyping, and evaluation. The literature review phase involved a targeted analysis of recent work on legal-domain LLMs, RAG, hallucination mitigation, and citation grounding. This provided the theoretical foundation for the research, for detailed background, see Section 3, *Related Work*.

4.1 Dataset Construction:

A curated set of 200 Dutch administrative rulings concerning social security disputes (e.g., disability benefits, unemployment benefits) were scraped from `Rechtspraak.nl`.

4.1.1 Temporal Stratification. Temporal bias can be introduced by collecting cases from a narrow time window, in order to avoid this a stratified sampling strategy was implemented to spread the dataset across five calendar years (2020-2024). The total case count was proportionally distributed across these years, and the rounding remainder was allocated to earlier years to preserve chronological balance. Each year's quota was further subdivided across its four quarters (Q1–Q4) to ensure that seasonal variation in rulings was captured. For each quarterly range, we queried the Rechtspraak API to retrieve paginated lists of ECLIs (European Case Law Identifier) [8] filtered by subject (`bestuursrecht_socialezekerheidsrecht`) and downloaded the corresponding XML documents.

4.1.2 Dataset Size Justification. Between 2020 and 2024, approximately 23,008 social security rulings were published. However, this project deliberately selected a subset of 200 cases due to methodological, computational, and practical considerations. Recent RAG literature highlights that retrieval over a smaller, high-quality, domain-focused corpus enhances citation accuracy and reduces hallucinations compared to using large, heterogeneous corpora [12, 15]. By limiting the dataset to 200 cases, the project ensures high topical relevance and avoids diluting retrieval quality with less pertinent or outlier rulings. Computationally, 200 rulings produce roughly 2,000 chunks, which is manageable for embedding, indexing, and evaluation within available resources, scaling up to 300 or 400 cases would significantly increase processing time without guaranteeing performance improvements. Moreover, preliminary sampling showed that the most common doctrines and procedural patterns are already well represented within the first few hundred rulings.

4.1.3 Dataset Limitations and Biases. `Rechtspraak.nl` is a curated platform, which means that only a selection of rulings that were deemed legally significant or instructive are published. This introduces a *publication bias*, potentially overrepresenting atypical or edge cases. Across the chosen legal category of social security cases, the distribution across legal subtopics is uneven because common disputes (e.g., WIA disability benefits) dominate the dataset while uncommon case types (e.g., AOW claims involving international law) are underrepresented. Thus, the real-world caseloads and editorial skew in what is published or not are reflected in the dataset. Moreover, court-level bias may affect retrieval relevance, while metadata such as `court` and `procedure` is preserved, the system does not yet differentiate between district rulings and higher court decisions, potentially grounding memos in less authoritative sources.

Lastly, linguistic variation across judges and jurisdictions can impact embedding-based similarity. Even with the `intfloat/multilingual-e5-large` model, semantic retrieval may favor more "typical" phrasings, skewing results toward dominant styles. These limitations underscore the importance of interpreting outputs with caution. Rich metadata per chunk enables future bias mitigation via filtering or reweighting.

4.1.4 Parsing and Structural Augmentation. The target documents were collected in their original XML format. XML is well-structured for legal publishing but not directly suitable for semantic processing, programmatic access, or integration with modern machine-learning pipelines due to its nested and verbose nature. This is why the 200 target documents were converted using the rechtspraak-js library [28, 29]. This tool was developed to extract semantically rich metadata from Dutch legal judgments, producing JSONL outputs compatible with linked data standards. Although highly effective at capturing metadata such as ECLI identifiers, court names, dates, and references, the parser deliberately excludes the full textual content of the judgment body in favor of metadata modeling.

Because of this limitation and to enable content-aware downstream applications, a secondary parsing step was introduced using the popular xml2js npm library [9]. This parser enabled deep traversal of the `<uitspraak>` element and extraction of the substantive textual content nested within `<section>`, `<paragroup>`, and `<parablock>` elements, which frequently contained critical legal reasoning and decisions rendered by the court.

The implemented dual-layer pipeline ensured comprehensive case representations by combining structured metadata with normalized full-text content. The paragraphs that were extracted were grouped under high-level legal sections such as *OVERWEGINGEN* (Considerations) and *BESLISSING* (Decision), and serialized under a unified `fullText` field. The final JSONL format facilitates retrieval, semantic search, and generative legal tasks that rely on both metadata accuracy and interpretability of legal arguments.

4.1.5 Chunking Strategy. After being parsed, the resulting JSONL files were segmented into context-preserving textual chunks to enable efficient retrieval within the RAG framework. A dual-threshold tokenization approach was implemented to ensure optimal chunk sizes for embedding generation and semantic search, which implied that short paragraphs (fewer than 50 words) were systematically

merged with adjacent content to prevent the creation of uninformative embeddings that could degrade retrieval performance.

The text segmentation process employed a hierarchical strategy that first split the content into sentences using regular expressions, then applied token-based boundary detection using the `intfloat/multilingual-e5-large` tokenizer. Each resulting chunk was constrained to contain between 50 and 512 tokens, ensuring compatibility with the embedding model's input requirements while maintaining sufficient semantic density. When paragraphs exceeded the maximum token limit, they were recursively split at sentence boundaries using a greedy accumulation algorithm that maximized chunk size while respecting the upper bound. In cases where individual sentences exceeded 512 tokens, controlled truncation was applied at the token level to preserve the most semantically relevant content. Chunks failing to meet the minimum 50-token threshold were either merged with adjacent content or accumulated with subsequent paragraphs until the threshold was satisfied. This approach eliminated fragmentary chunks that could introduce noise into the vector space while preserving document coherence. All resulting chunks were enriched with comprehensive metadata including ECLI identifier, section classification, document title, issuing court, judgment date, legal subject matter, and procedural type to enable precise filtering and contextualization during retrieval. A detailed validation of chunk completeness, token length distribution, and quarterly representativeness is provided in Appendix A.3.1. Each abstract (when available in JSON records) was included as a separate chunk with a designated index of -1 in order to provide contextual grounding for generations.

4.1.6 Embedding and Indexing. Textual chunks were embedded using the `multilingual-e5-large` model and stored in a pgvector-augmented PostgreSQL database hosted on Supabase. A multi-tiered indexing strategy was employed to enable fast semantic retrieval and structured filtering. For implementation details, see Appendix A.3.2.

4.2 System Prototyping.

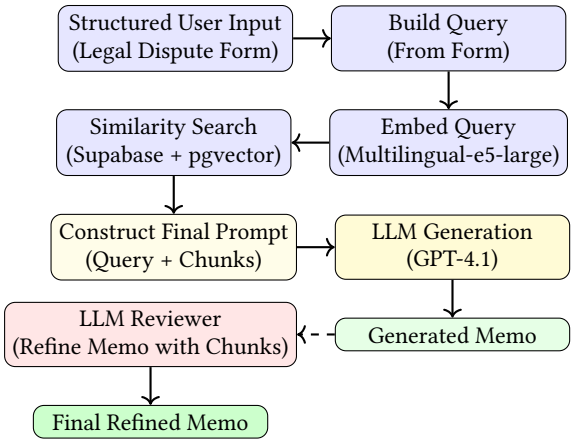


Fig. 1. RAG pipeline with optional LLM reviewer refinement

4.2.1 Structured User Input and Query Building. To support grounded and personalized memo generation, the system prompts the user to complete a structured legal intake form. This form captures five key aspects of an administrative legal dispute, detailed in Table 1 in Appendix A.4. These user inputs are the semantic foundation for query formation and later chunk retrieval. This structure allows the system to minimize ambiguity and improve interpretability. It also follows recommendations by Murray [20], who emphasizes the importance of context-rich prompting that includes the legal role, relevant norms, and procedural setting to improve LLM fidelity in legal tasks. An example of completed input fields is included in Appendix Example Structured Input.

4.2.2 Similarity Search. A custom `supabase` remote procedure called `match_case_chunks` [27] was implemented using the `pgvector` extension [14]. This enables accurate and efficient retrieval over the embedded case law corpus. The remote function initially retrieves up to 50 candidate chunks to ensure sufficient diversity before applying the per-ECLI filtering constraints, ultimately returning the top-k most relevant results based on a minimum similarity threshold of 0.70. An example of the structure of the retrieved chunk object can be found in Appendix Example Retrieved Chunk Structure.

Each retrieved chunk is augmented with its globally unique identifier as stored in the `supabase` database, along with fine-grained indices corresponding to its section origin and paragraph position, this object design enables precise mapping between the retrieved chunks and downstream feedback mechanisms, including chunk-specific relevance judgments and user comments. These identifiers are also important when auditing retrieval behavior in experiments and production settings. To mitigate the risk of retrieval redundancy, where multiple top-k results originate from the same legal ruling, a grouping strategy was implemented at the application level. This implies that when retrieved chunks are first grouped by their ECLI, only a limited number of chunks per ruling (maximum two) are retained. This approach promotes diversity across retrieved legal sources, ensuring the generated memo benefits from a broader legal context rather than repetitive evidence from a single case.

4.2.3 Prompt Construction LLM as Generator Step. The prompt for the first LLM generation step consists of four key components: role priming, the structured user query, the retrieved case law fragments, and explicit instructions for memo synthesis.

The prompt first assigns the model a professional identity: "Je bent een juridisch assistent gespecialiseerd in Nederlandse sociale zekerheidszaken." ("You are a legal assistant specialized in Dutch social security cases."). This follows the role-specific priming strategy mentioned by Kang et al. [13], who showed that assigning a legal role improves the fidelity of reasoning and consistency of output when prompting ChatGPT to follow the IRAC style (Issue, Rule, Application, Conclusion) for legal reasoning.

Strict context containment is enforced by instructing the model to rely exclusively on the retrieved jurisprudence and to cite the relevant ECLI in every sentence where the precedent is used. This aligns with hallucination prevention techniques from Ioannidis et al. [11] and Schwarcz et al. [26], both of whom demonstrate that hallucination rates can be minimized by grounding generation in retrieved source material and by making citations mandatory. Furthermore, Posner & Saran [23] found that human readers rated LLM-generated legal arguments as more credible when citations were consistent and easy to trace.

A structured three-part format is explicitly requested from the model for each of the legal memos: Vraaganalyse (Issue Analysis), Toepassing jurisprudentie (Application of Case Law), and Conclusie

(Conclusion) which mirrors the IRAC-style scaffolding shown to improve legal reasoning coherence in Kang et al. (2023) [13] and Martinez [19]. Stepwise and sectional prompts were consistently associated with better-structured and more accurate responses.

4.2.4 Reviewer LLM as a Post-Generation Alignment Step. To improve the legal factuality and citation grounding beyond first-pass generation, a secondary LLM step dedicated to post hoc memo refinement was introduced. After the initial memo is generated with the structured legal query and the set of retrieved jurisprudence chunks, the output is forwarded along with the same chunk set and metadata to a second LLM invocation. The model is tasked with reviewing the memo's alignment with the retrieved content and rewriting it to remove unsupported claims, reinforce citations, and improve legal clarity.

4.2.5 Prompt Construction LLM as Reviewer Step. In the reviewer step, the prompt mirrors the structure and role-based priming of the generation phase but shifts focus from creation to correction. The model is given the same professional identity, "Je bent een juridisch assistent gespecialiseerd in Nederlandse sociale zekerheidszaken" but is now instructed to assess the legal accuracy and grounding of a previously generated memo using the retrieved court decisions. The input includes both the memo and source fragments, and the model is tasked with removing unfounded claims, correcting or adding ECLI citations, and refining the legal reasoning to enhance clarity and trust.

4.3 Explainability and Transparency in Legal AI Systems

Legal professionals must not only evaluate the conclusions of AI-generated outputs but also understand the reasoning behind them and assess their evidentiary basis, that is why explainability is a critical requirement for AI applications in law. During this research, explainability and transparency were core design principles that were implemented in both the back-end and front-end layers.

To ensure that legal professionals articulate their assumptions before relying on the generated output, the system begins with a structured intake form that prompts users to specify five key dimensions of the legal dispute: the contested decision, the desired legal outcome, relevant factual context, applicable legal norm, and the memo's intended audience (Appendix B.1 Figure 7). After generation, the interface displays the legal sources used during memo generation alongside the output, with each legal fragment shown together with its ECLI, similarity score, and excerpted content (Appendix B.1 Figure 10). This allows users to verify each claim made in the generated memo against the source text.

To support transparency and iterative review, users are encouraged to mark each retrieved fragment as "Relevant" or "Not Relevant" and provide short justifications (Appendix B.1 Figure 9). They can also leave general feedback on the clarity, structure, and persuasiveness of the memo, positioning themselves as active reviewers rather than passive consumers (Appendix B.1 Figure 9 and Appendix B.1 Figure 8). This user interaction fosters critical engagement and supports the principle of *procedural transparency*: users know what information the system received, how it processed that information, and which sources were used in its reasoning. This approach goes beyond post hoc interpretability, instead making the generation process itself auditable and aligned with legal expectations for traceable reasoning. Additional interface components, such as the disclaimer presented before generation (Appendix B.1 Figure 8) and the searchable jurisprudence database used by the LLM in the generation step (Appendix B.1 Figure 11), further promote awareness and control over the system's behavior and knowledge base.

Overall, explainability in legal AI should prioritize verifiability, traceability, and user agency. By requiring structured input, enforcing transparent citation, and enabling legal users to scrutinize source material, the system shows that these goals can be embedded not only in model design but directly into the interface layer.

4.4 Evaluation

A modular evaluation pipeline was designed to assess the factual reliability, legal grounding, and hallucination risk of the generated legal memos, the pipeline combines automated information retrieval metrics, semantic similarity comparisons, and optional human or LLM-based verification. The evaluation extracts all ECLI citations from a memo using a regular expression to produce a list of predicted citations, which are then compared to the reference citations found in the retrieved jurisprudence chunks. Two metrics are computed: citation precision (the fraction of predicted ECLIs present in the retrieved context) and citation recall (the fraction of retrieved ECLIs that are actually cited in the memo). Together, these metrics are used to compute an IR-style (Information Retrieval) F1 score, enabling the distinction between over-citation (hallucinated references) and undercitation (omitted references).

In addition to citation fidelity, semantic grounding at the sentence level is assessed. Each sentence in the memo is embedded using the `multilingual-e5-large` model and compared to all retrieved chunk embeddings using a configurable similarity metric (cosine, dot-product, or Euclidean distance). When the similarity score is below a predefined threshold (e.g. 0.7) the sentence is flagged as *ungrounded*. The hallucination flag for a memo is set to true if either fabricated citations or ungrounded sentences are detected. Each evaluation produces a structured output that can be found in Appendix: Example Evaluation Object Structure.

The pipeline can be run over a grid of similarity metrics and different thresholds, which facilitates ablation studies and metric robustness checks. Furthermore, all evaluation outputs, along with the corresponding memo and retrieved chunks, are saved to a dedicated Supabase table for tracking and reproducibility. Finally, an LLM also checks a sample of evaluation results to get a different perspective on what is classified as factually unsupported or supported. This allows for comparison between the similarity-based heuristics and a reasoning-based proxy, quantifying agreement rates and analyzing false positive or false negative patterns.

4.4.1 Interpretation of Expert Feedback. Two practicing legal professionals offered their feedback regarding the live memo generator system. Both reviewers appreciated the structured layout and found that the memo output followed a format familiar to legal work, combining case description, legal analysis, and conclusion. They also highlighted areas for improvement, particularly in refining legal nuance and ensuring clearer distinction between factual and normative reasoning. However, this expert feedback should be interpreted as indicative rather than conclusive, as it was based on a small sample of two reviewers and arrived relatively late in the project. Rather than being part of the formal evaluation phase, it was used to inform fine-tuning decisions during system refinement.

As such, it serves more as a qualitative perspective than a rigorous evaluation benchmark.

One reviewer noted that the summaries of the cited case law were not always clearly aligned with the facts of the specific case, which points out a need for greater contextual specificity. Chunk-level relevance judgments reinforced the same observation: none of the retrieved jurisprudence fragments in one case were deemed directly relevant to the core legal issue (Appendix C.1 Table 9). This puts forward a need for retrieval strategies that go beyond surface similarity and account for the procedural stage, statutory domain, and legal framing. Another recurring theme was the desire to better understand how the AI system arrived at its conclusions. This makes the future addition of rationale generation a logical step to ensure more transparent and interpretable memo generation. Reviewers also valued the ability to group and compare multiple similar rulings, which shows that the system's structure aligns well with legal research patterns.

The system's core design choice to expose legal sources and allow user review was validated by the feedback, but the legal professionals also stressed the *importance of semantic retrieval precision, interpretive clarity, and role-sensitive explanation* for real-world legal deployment. The two appended tables summarize the general and chunk-specific feedback from these professionals (see Appendix C.1 Table 8 and Table 9).

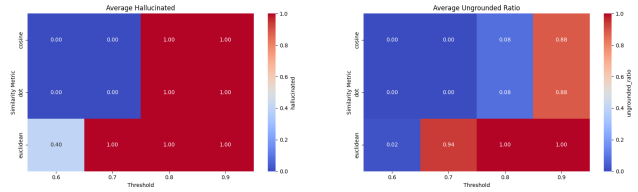
5 EXPERIMENTS

First, the *Baseline RAG System*, representing the standard RAG pipeline without any additional reviewer step was evaluated, then the *Enhanced RAG System*, which augmented the baseline with a post hoc LLM reviewer that examined generated content for factual accuracy and source grounding. For a comprehensive evaluation, a parameter grid was constructed by sweeping systematically across multiple dimensions. Three similarity metrics (cosine similarity, dot product, and Euclidean distance), four grounding thresholds (0.6, 0.7, 0.8, and 0.9), and 15 representative Dutch administrative court cases composed the parameter space. This created a $3 \times 4 \times 15$ evaluation matrix, yielding 180 individual assessments per system configuration and a total of 360 evaluations.

Two sequential phases composed the evaluation protocol, in Phase 1 the baseline RAG system was run across all 15 cases, and for each generated memo, citation and grounding metrics were computed across all 12 parameter combinations. Phase 2 used the same procedure but with an enhanced system, which allowed for direct comparison between configurations. For each experimental run, the generated memo, retrieved chunk metadata, citation precision, recall, and F1 scores, as well as sentence-level grounding evaluations and hallucination indicators such as ungrounded ratios and fabricated citations were saved for later analysis. Furthermore, all outputs were logged with timestamps, script versions, and git commit hashes to ensure full reproducibility and traceability over time.

In the final step, GPT-4.1 was used to review evaluation outputs and provide an alternative perspective on which similarity-based predictions were ungrounded. This was applied to both baseline and enhanced systems, enabling multi-dimensional analysis, performance comparison, sensitivity testing, and robustness checks across evaluation metrics.

5.1 Baseline Memo Generation



(a) Hallucinated heatmap - without reviewer LLM

(b) Ungrounded ratio heatmap - without reviewer LLM

Fig. 2. Heatmaps of hallucination (a) and ungrounded ratio (b) across memo sections without reviewer LLM. Each cell shows the average error for a section–threshold pair, with red indicating higher error rates. Errors increase at higher thresholds.

The citation precision demonstrated by the system was perfect at 1.0, indicating that no fabricated ECLIs were present in any generated memo. However, citation recall remained consistently at 0.78, suggesting that while all cited ECLIs were valid, a portion of the retrieved sources were not explicitly cited in the final output. The F1 score, which is the harmonic mean of precision and recall, remained stable at 0.87 across all configurations, highlighting the system's robust citation behavior.

The hallucination behavior concerning semantic grounding showed more variability across similarity thresholds. At low thresholds (0.6-0.7), the model produced almost no ungrounded sentences, resulting in an ungrounded ratio of 0.0 for dot and cosine metrics and approximately 0.4 for Euclidean distance. However, as similarity thresholds increased to 0.8 and 0.9, the evaluation flagged ungrounded statements with increasing frequency. This rise was especially sharp for Euclidean distance, where the ungrounded ratio climbed from approximately 0.4 at threshold 0.6 to 1.0 (complete ungrounding) at thresholds 0.7 and above.

The similarity metrics chosen proved consequential for grounding evaluations as cosine and dot product similarities exhibited identical behavior patterns, by maintaining zero ungrounded statements until threshold 0.8, where the ungrounded ratio jumped to 0.08, and further escalated to 0.88 at threshold 0.9. In contrast, euclidean distance demonstrated greater sensitivity to threshold changes, with the ungrounded ratio climbing from 0.02 at threshold 0.6 to 0.94 at threshold 0.7, ultimately reaching complete ungrounding (ratio of 1.0) at higher thresholds.

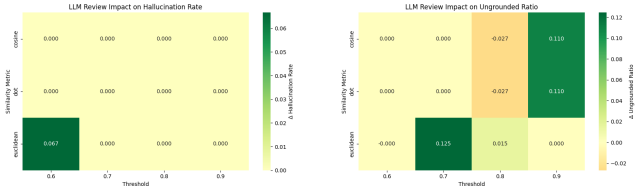
In conclusion, the results suggest that the baseline system demonstrated strong citation accuracy but faced challenges in maintaining semantic coherence with source materials when evaluated by stricter similarity criteria. A threshold of 0.6-0.7 with cosine or dot product similarity appears to represent a balanced choice between evaluation precision and realistic assessment of content fidelity.

5.2 Memo Generation with Reviewer LLM

The post hoc LLM review integration into the RAG pipeline yielded disappointingly modest improvements in content grounding while maintaining the system's existing citation accuracy. The enhanced system preserved perfect citation precision (1.0) across all configurations, but this doesn't represent an advancement over the baseline system, which already achieved zero fabricated ECLI citations. The citation recall remained stagnant at 0.78, identical to the baseline

system, indicating that the reviewer component did not enhance the system's ability to comprehensively cite available sources. The small improvements in semantic grounding performance proved to be mostly illusory when examined closely. At optimal thresholds (0.6, 0.7), both systems achieved identical ungrounded ratios of 0.0 for cosine and dot product metrics, rendering the reviewer component mostly redundant at these operating points. At higher thresholds, some marginal differences emerged at inconsistent rates, suggesting limited practical value from the additional computational overhead.

At threshold 0.8, the enhanced system's performance was paradoxically worse than the baseline for cosine and dot product metrics, with ungrounded ratios increasing to 0.11 compared to the baseline's 0.08, suggesting that the reviewer component may have introduced noise or inconsistency into the grounding evaluation process. Finally, at the most restrictive threshold of 0.9, the ungrounded ratio decreased to 0.77 compared to the previous 0.88. However, this improvement occurred only at the most stringent threshold, making it practically irrelevant for real-world deployment.



(a) Change in hallucination scores after reviewer LLM intervention.

(b) Change in ungrounded ratio after reviewer LLM intervention.

Fig. 3. Delta heatmaps showing changes in hallucination (a) and ungrounded ratio (b) after reviewer LLM intervention. Green indicates error reduction; orange shows increases. Results reveal mixed effects and modest gains at some thresholds, regressions at others.

The Euclidean distance metric made the reviewer's limited effectiveness more evident as at threshold 0.6, there was no difference in ungrounded ratios, then at threshold 0.7 while the ungrounded ratio of 0.81 appeared better than the baseline's of 0.94, this still represented a substantial failure of over 80% ungrounded content. The marginal improvement at threshold 0.8 (0.99 vs. 1.0 baseline) constituted a practically meaningless distinction between near-complete and complete ungrounding.

Even though perfect hallucination prevention (0.0) was maintained at thresholds 0.6-0.7 for cosine and dot product metrics, this just replicated the baseline system's already good performance. Moreover, there was only a marginal reduction in Euclidean hallucination rates (0.33 vs 0.4 baseline at threshold 0.6) which can fall within measurement error. At higher thresholds, the binary hallucination pattern (1.0) remained unchanged, indicating that the reviewer failed to prevent the fundamental grounding failures that plague the system under stricter evaluation criteria.

The F1 score's stability at 0.87 across all configurations exposes the inability of the reviewer component to meaningfully enhance citation behavior, which is the most critical aspect of legal memo reliability. This stagnation confirms that the enhanced system provided no tangible benefit for the primary use case of accurate legal reference generation.

These results reveal that post hoc LLM review constitutes an inefficient allocation of computational resources, providing marginal and inconsistent improvements that fail to justify the additional processing overhead.

5.3 Review of Ungrounded Labeled Statements by an LLM

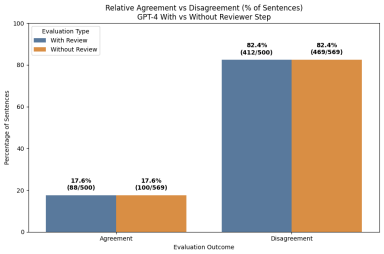


Fig. 4. Relative agreement and disagreement rates between GPT-4 evaluations with and without the reviewer LLM. Bars show both percentage and count of sentence-level judgments, normalized by the number of sentences per condition for fair comparison.

To get a different perspective on the automated grounding heuristics, each sentence initially flagged as ungrounded by the heuristic evaluation was re-evaluated by GPT-4.1 using a structured prompt to determine whether it was hallucinated or just a false positive. The analysis compared both the baseline system and the reviewer-enhanced system across 30 evaluation cases. The overall agreement between heuristic flags and GPT verdicts remained low despite the addition of a reviewer step intended to improve factual consistency. Only 10 out of 30 cases showed majority agreement. These were cases with no sentences marked as ungrounded due to low thresholds (0.6, 0.7). This yielded a stable agreement rate of 33% across both configurations.

Sentence-level verdicts revealed consistent disagreement patterns. For instance, in test-case-3 [27], the automated evaluation system flagged 29 sentences generated by the baseline system as ungrounded, but GPT-4.1 judged that only 11 of these were actual hallucinations. In comparison, the enhanced system flagged slightly fewer sentences (25), with GPT-4.1 judging that only 3 of these were actual hallucinations.

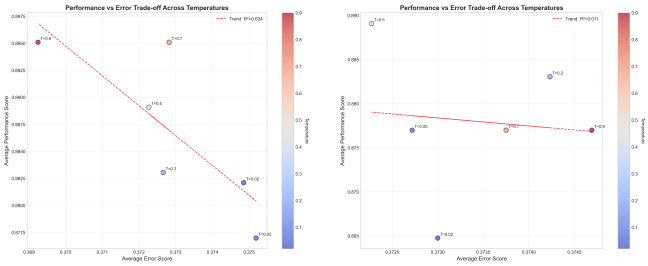
A similar pattern occurred in test-case-5 [27], where 31 sentences were flagged in the baseline, and only two were confirmed as hallucinated by the LLM for the enhanced system, with the number of flagged sentences decreasing modestly to 21 while GPT-4.1 marked 6 of those as hallucinated. In all reviewed cases, the reviewer step led to changes in sentence counts but did not alter the overall GPT majority verdict, indicating that these differences were insufficient to flip the final judgment.

The LLM's hallucination detections remained largely stable in both baseline and enhanced system scenarios, the verdict comparison summary confirming that in all 30 evaluation cases, the GPT majority classifications were unchanged. No cases were "resolved" (i.e. changed completely to non-hallucinated (0)) or "regressed" (i.e. to hallucinated (1)). This suggests that while the reviewer may influence local sentence phrasing or structure, it does not substantially improve the legal factuality or grounding traceability in a way that GPT-4.1 can detect.

These findings highlight a disconnect between heuristic thresholds, especially at high similarity cutoffs, and the LLM's legal reasoning assessments. Despite reducing the number of flagged sentences

in some cases, the reviewer component did not meaningfully impact the LLM's evaluation outcomes.

5.4 Temperature Changes for Reviewer LLM



(a) Performance–error trade-off for GPT-4.1 across temperature settings.

(b) Performance–error trade-off for Claude Sonnet 4 across temperature settings.

Fig. 5. Performance–error trade-off curves for GPT-4.1 and Claude Sonnet 4 across temperature settings ($T = 0.02$–0.9). Each plot shows memo quality (citation precision, recall, F1) versus grounding errors (hallucination rate, ungrounded ratio) over 1000+ runs. GPT-4.1 shows a positive trade-off, higher temperatures improve quality with limited error increase while Claude remains largely unaffected by temperature changes.

The impact of temperature on the behavior and effectiveness of the reviewer LLM was tested through a controlled series of experiments. Two state-of-the-art LLMs were used: GPT-4.1 and Claude Sonnet 4. Each model was evaluated across a range of temperature settings, from near-deterministic outputs at $T = 0.02$ to highly generative behavior at $T = 0.9$. The experiment used the same 15 Dutch administrative court cases and followed the same evaluation procedure as in previous sections, computing grounding metrics for each refined memo across all combinations of similarity metrics (cosine, dot product, Euclidean) and grounding thresholds (0.6-0.9). This resulted in over 1000 unique evaluation runs per model, allowing for a detailed comparative analysis of how temperature affects reviewer consistency and overall memo quality across different levels of LLM creativity.

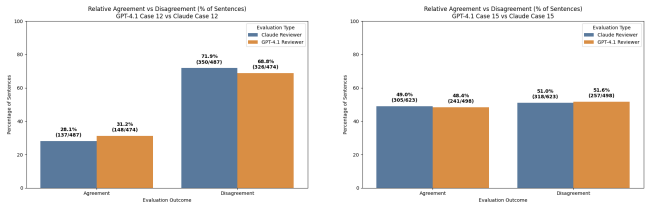
5.4.1 GPT-4.1 As Reviewer. Temperature tuning had a measurable but modest effect on grounding-related evaluation metrics for the GPT-4.1 reviewer, keeping the citation metrics almost unchanged across all temperatures from $T = 0.02$ up to $T = 0.9$. The reviewer consistently achieved perfect citation precision (1.0) and showed stable citation recall between 0.76 and 0.80 which resulted in a highly consistent F1 score ranging from 0.87 to 0.89, confirming that altering the temperature does not meaningfully affect the reviewer's ability to maintain citation correctness. Ungrounded ratios and hallucination rates remained negligible for cosine and dot metrics at thresholds 0.6 and 0.7, with values as low as 0.0, but as temperature increased, these metrics worsened modestly, particularly at threshold 0.8, where the ungrounded ratio rose to approximately 0.10-0.11 and hallucination rates increased to 0.87-0.93.

A performance error trade-off analysis (see Figure 5a) revealed that higher temperatures, especially $T = 0.7$ and $T = 0.9$, offered the best balance between strong citation behavior and grounding errors. Furthermore, a clear negative correlation between performance score and error score is seen in the trade-off curve, with a trend line slope indicating that increases in temperature improved citation completeness and F1 scores more than they degraded citation grounding. In contrast, very low temperatures ($T = 0.02$ and $T = 0.05$) led to slightly lower performance, even though hallucination rates remained low. This suggests that excessively deterministic reviewers may under-refine memos. This trend, visualized by the regression fit of $R^2 = 0.634$, suggests that a slight increase in reviewer creativity can make legal memos better, but temperature is still a secondary optimization factor rather than the primary driver of reviewer effectiveness.

5.4.2 Claude Sonnet 4 As Reviewer. Temperature adjustments yielded similarly stable citation behavior for the Claude Sonnet 4 reviewer but with more pronounced variability in semantic grounding metrics compared to GPT-4.1. Across all temperature levels ($T = 0.02$ to $T = 0.9$), the model consistently achieved perfect citation precision (1.0) and maintained a nearly constant citation recall of approximately 0.74 to 0.76 which resulted in a narrow F1 score kept between 0.85 and 0.88, indicating highly consistent citation coverage regardless of generation randomness.

Semantic grounding measures, including hallucination rate and ungrounded ratio, responded more noticeably to temperature changes. At $T = 0.02$ to $T = 0.2$, the cosine and dot product similarity metrics flagged virtually no ungrounded sentences at thresholds 0.6 and 0.7, and hallucination rates remained at 0.0. However, as temperature rose to $T = 0.5$ and $T = 0.7$, these metrics deteriorated at threshold 0.8, with ungrounded ratios increasing to around 0.10 and hallucination rates to 0.9. Euclidean similarity again proved more sensitive and less reliable across all temperature settings because even at the lowest threshold of 0.6, ungrounded ratios for Euclidean distance exceeded 0.65, and hallucination rates were consistently flagged as 1.0 regardless of temperature.

No meaningful correlation between temperature and overall system effectiveness was seen in the performance-–error trade-off curve for Claude (Figure 5b). Although $T = 0.5$ stood out as a slight high point in performance, the overall linear trend across temperatures was almost flat with an $R^2 = 0.011$, suggesting that the reviewer's average performance and error scores were largely unaffected by changes in generative randomness. Claude's outputs appeared to be more stable and indifferent to temperature compared to GPT-4.1, reinforcing the conclusion that temperature tuning has limited utility for post hoc legal memo refinement.



(a) Comparison of agreement for case 12 with reviewer as GPT-4.1 and Claude Sonnet 4.

(b) Comparison of agreement for case 15 with reviewer as GPT-4.1 and Claude Sonnet 4.

Fig. 6. Sentence-level comparison of agreement and disagreement between similarity-based hallucination flags and GPT-4.1's structured verdicts, using GPT-4.1 and Claude Sonnet 4 as independent reviewers. The figure highlights how reviewer model choice influences alignment with heuristic flags, revealing both consistent trends and case-specific differences in hallucination detection.

5.4.3 Review of Ungrounded Labeled Sentences by an LLM for GPT-4.1 and Claude Sonnet 4 Temperature Experiments. In order to get a different perspective into how reviewer models affect the heuristic hallucination detection, Figure 6 compares the sentence-level agreement between similarity-based flags and GPT-4.1's verdicts for both GPT-4.1 and Claude Sonnet 4 as reviewers. Firstly, for case 12, the agreement with GPT-4.1 was low for both reviewers, but lower for Claude (28.1%) compared to GPT-4.1 (31.2%). Disagreements dominated for both, suggesting that many flagged sentences may be paraphrased or legally defensible rather than truly hallucinated. Secondly, for case 15, both models showed nearly identical agreement rates: 49.0% for Claude and 48.4% for GPT-4.1. The remaining half of sentences were marked as hallucinations not confirmed by GPT-4.1. A consistent pattern is highlighted by these results: both reviewer-enhanced systems still over-flag valid sentences at higher thresholds, and neither GPT-4.1 nor Claude is better at avoiding false positives when judged by a second-pass LLM. The near-identical disagreement rates suggest that the choice of a state-of-the-art reviewer model has a limited impact on the general alignment with the GPT-4.1 groundedness assessment, reaffirming the limitations of heuristic similarity cutoffs regardless of reviewer architecture.

6 DISCUSSION

6.1 Interpretation of Core Findings

Several key insights emerged during the experimentation phase. The perfect citation precision (1.0) across all system configurations demonstrates that properly designed RAG architecture can eliminate fabricated references, which addresses one of the most serious ethical concerns regarding explainability in AI deployment for legal practices. It suggests that when retrieval mechanisms are robust and promptly enforce strict source containment, LLMs can reliably avoid generating nonexistent case citations that may lead to wrong information influencing practitioners or courts.

However, the consistent citation recall of 0.78 indicates systematic undercitation of available precedent. This suggests that while the system can successfully avoid fabrication, it may miss opportunities to cite relevant authority that could strengthen legal arguments. The failure of post-generation LLM reviewers to meaningfully improve performance challenges prevalent assumptions about multi-step verification in legal AI. Despite significant computational overhead, the reviewer component provided only marginal and inconsistent improvements, sometimes even degrading semantic coherence. This suggests that current approaches to mitigate hallucination through architectural complexity may be misguided for state-of-the-art LLMs in 2025.

6.2 Evaluation Framework Limitations and Implications

A critical limitation in automated legal AI evaluation was exposed by the substantial disagreement between similarity-based hallucination detection and expert LLM judgment (33-69% agreement rates), with high similarity thresholds frequently flagging legitimate legal reasoning as hallucinated. This disconnect has immediate practical implications: legal professionals cannot rely solely on automated safeguards to assess AI-generated content quality and must review each statement for legal grounding themselves, which might defeat the argument for increased efficiency. The choice of similarity metric significantly impacts evaluation outcomes, with cosine similarity proving more aligned with legal reasoning patterns than geometric distance measures. Euclidean distance metrics performed poorly compared to cosine similarity, which provides an actionable precedent for practitioners implementing similar systems.

6.3 Practical Deployment Considerations

The system was limited to 200 social security cases within a specialized legal domain, inheriting the publication bias from `Rechtspraak.nl`. This may have benefited system performance by focusing only on high-quality precedent, though it limits representation of routine legal reasoning patterns. Future practitioners should keep the retrieval database as small and focused as possible on the niche system that they want to implement to ensure optimal resource utilization.

However, in the case of the legal memo generator, the inability to distinguish between hierarchical court authorities represents a significant limitation for real-world deployment because legal practice requires an understanding that district court rulings carry less precedential weight than appellate decisions. Yet, the current system treats all retrieved sources equally. Future implementations must incorporate metadata-based authority weighting to align with established legal hierarchy principles.

7 CONCLUSION

By developing and evaluating a RAG system specifically designed for Dutch legal memo generation, this research aims to address a critical gap in current research by emphasizing citation reliability and hallucination control. The experiments demonstrated that well-designed retrieval mechanisms that take full advantage of prompt engineering can achieve perfect citation precision while maintaining acceptable semantic grounding, providing a viable foundation for legal AI deployment.

This thesis made three primary contributions to legal AI research. First the set-up of a systematic evaluation of RAG for Dutch legal memo generation which established baseline performance metrics for citation fidelity and semantic grounding in legal contexts. The second contribution is a novel evaluation framework that integrates automated citation verification with similarity-based grounding analysis, augmented by an additional LLM-based agreement layer that independently assesses whether flagged ungrounded statements truly constitute hallucinations. Third, and most significant, is the demonstration that expensive post-generation verification layers provide minimal benefits when upstream components are properly designed.

The effectiveness of legal AI requires prioritizing data quality and retrieval precision over architectural complexity. This insight provides actionable guidance for legal technology developers as computational resources should be invested in corpus curation based on legal cases, chunking strategies, and retrieval optimization rather than multistep verification systems. This finding can help to simplify future RAG architecture while not missing out on any added value. The expansion of the legal corpus beyond social security law, the incorporation of hierarchical authority weighting, and the development of domain-specific evaluation metrics should be the focus of future research as they can help the system better align with legal reasoning patterns. This evaluation framework supports scalable legal AI while emphasizing that trustworthy systems come from better inputs, not more layers.

REFERENCES

[1] Vinayshekhar Bannihatti Kumar, Kasturi Bhattacharjee, and Rashmi Gangadharaiah. 2022. Towards Cross-Domain Transferability of Text Generation Models for Legal Text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, and Daniel Preoțiuc-Pietro (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 111–118. https://doi.org/10.18653/v1/2022.nllp-1.9

[2] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of law school. In *Findings of EMNLP*. 2898–2904. https://aclanthology.org/2020.findings-emnlp.261/

[3] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16, 1 (2024), 64–102. https://arxiv.org/html/2401.01301v1

[4] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv:1912.09582 [cs.CL] https://arxiv.org/abs/1912.09582

[5] Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 3255–3265. https://doi.org/10.18653/v1/2020.findings-emnlp.292

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[7] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-Verification Reduces Hallucination in Large Language Models. arXiv:2309.11495 [cs.CL] https://arxiv.org/abs/2309.11495

[8] EUR-Lex. 2025. European Case Law Identifier (ECLI). https://eur-lex.europa.eu/content/help/eurlex-content/ecli.html Accessed: 2025-04-29.

[9] Leonidas from XIV. 2023. xml2js: Simple XML to JavaScript object converter. https://www.npmjs.com/package/xml2js. Accessed: 2025-05-21.

[10] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. https://doi.org/10.1145/3703155

[11] Jules Ioannidis, Joshua Harper, Ming Sheng Quah, and Dan Hunter. 2023. Gracenote.ai: Legal Generative AI for Regulatory Compliance. In *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023)*. https://dx.doi.org/10.2139/ssrn.4494272

[12] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. arXiv:2112.09118 [cs.IR] https://arxiv.org/abs/2112.09118

[13] Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Zhuo, Patrick Emerton, and Genevieve Grant. 2023. Can ChatGPT Perform Reasoning Using the IRAC Method in Analyzing Legal Scenarios Like a Lawyer?. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13900–13923. https://doi.org/10.18653/v1/2023.findings-emnlp.929

[14] Jonathan Katz. 2023. pgvector: Open-source vector similarity search for PostgreSQL. https://github.com/pgvector/pgvector. Accessed May 2025.

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/2005.11401

[16] M.S. Looijenga. 2024. RechtBERT : Training a Dutch Legal BERT Model to Enhance LegalTech. http://essay.utwente.nl/104811/

[17] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems (NeurIPS) 36*. https://papers.nips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf

[18] Varun Magesh, Faiz Surani, Mirac Suzgun, Matthew Dahl, and Daniel E. Ho. 2025. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies* (2025), 1–27. https://doi.org/10.1111/jels.12413

[19] Eric Martínez. 2024. Re-evaluating GPT-4's Bar Exam Performance. *Artificial Intelligence and Law* (2024). https://doi.org/10.1007/s10506-024-09396-9 Published: March 30, 2024.

[20] Michael D. Murray. 2024. Prompt Engineering and Priming in Law. https://dx.doi.org/10.2139/ssrn.4909532. SSRN preprint.

[21] Minhu Park, Hongseok Oh, Eunkyung Choi, and Wonseok Hwang. 2025. LRAGE: Legal Retrieval Augmented Generation Evaluation Tool. arXiv:2504.01840 [cs.CL] https://arxiv.org/abs/2504.01840

[22] Nicholas Pipitone and Ghita Houir Alami. 2024. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. arXiv:2408.10343 [cs.AI] https://arxiv.org/abs/2408.10343

[23] Eric A. Posner, Shivam Saran, and Chicago Law RPS Submitter. 2025. *Judge AI: Assessing Large Language Models in Judicial Decision-Making*. Technical Report. University of Chicago Coase-Sandor Institute for Law & Economics Research Paper No. 25-03. https://ssrn.com/abstract=5098708

[24] Felicia Redelaar, Romy Van Drie, Suzan Verberne, and Maaike De Boer. 2024. Attributed Question Answering for Preconditions in the Dutch Law. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, Daniel Preoțiuc-Pietro, and Gerasimos Spanakis (Eds.). Association for Computational Linguistics, Miami, FL, USA, 154–165. https://doi.org/10.18653/v1/2024.nllp-1.12

[25] Marijn Schraagen, Floris Bex, Nick Van De Luijtgaarden, and Daniël Prijs. 2022. Abstractive Summarization of Dutch Court Verdicts Using Sequence-to-sequence Models. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, and Daniel Preoțiuc-Pietro (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 76–87. https://doi.org/10.18653/v1/2022.nllp-1.7

[26] Daniel Schwarcz, Sam Manning, Patrick James Barry, David R. Cleveland, J.J. Prescott, and Beverly Rich. 2025. *AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice*. Technical Report. Minnesota Legal Studies Research Paper No. 25-16. https://ssrn.com/abstract=5162111

[27] Mihai Timoficiuc. 2025. Legal Memorandum Generation Using Retrieval-Augmented Large Language Models and Dutch Case Law. https://github.com/MTimo27/rag-dutch-law-memo-generator.

[28] M.F.A. Trompper. 2016. *Automatic Assignment of Section Structure to Texts of Dutch Court Judgments*. Master Thesis. Utrecht University, Utrecht, The Netherlands. https://studenttheses.uu.nl/handle/20.500.12932/24346 Advised by A. Feelders. Keywords: Conditional Random Fields; Probabilistic Context-Free Grammars; automatic markup; court judgments.

[29] M.F.A. Trompper. 2016. rechtspraak-js: Tools for parsing and handling Rechtspraak.nl XML data. https://github.com/digitalheir/rechtspraak-js. GitHub repository accompanying the master thesis on automatic section structure assignment.

[30] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. arXiv:2402.05672 [cs.CL] https://arxiv.org/abs/2402.05672

A APPENDIX A

A.1 AI tool use

During the preparation of this thesis, the author used ChatGPT and Grammarly to assist with proofreading and improving the structure of the document. All original content was created by the author. The tools were used solely for language refinement and structural suggestions. Final responsibility for the content rests fully with the author.

A.2 Related Work

To gather relevant literature in the research domain, databases such as Google Scholar, arXiv, and ACL Anthology were consulted. Search terms included *"Retrieval-Augmented Generation"*, *"legal NLP"*, *"hallucination reduction"*, *"citation grounding in LLMs"*, and *"Dutch case law AI"*.

There has been a growing interest in the development of domain-adapted language models for legal NLP. One of the earliest and most influential models is Legal-BERT[2], which demonstrated that pre-training or further training on legal corpora significantly improves performance on legal tasks such as statute classification, legal topic tagging, and named entity recognition. Following this approach in the Dutch context, RechtBERT[16] was developed by further pre-training three Dutch BERT variants (BERTje[4], RobBERT[5], and mBERT [6]) on over 450,000 Dutch court rulings from `Rechtspraak.nl` and statutory texts from `wetten.overheid.nl`. Although RechtBERT did not consistently outperform the base models on EU legislation classification, it confirmed that legal-domain fine-tuning is viable and may offer performance gains in more specialized downstream tasks such as document analysis and legal search. However, these models are still limited by fixed input lengths, domain-specific corpus availability, and lack of grounding mechanisms.

Several studies have highlighted the benefits of Retrieval- Augmented Generation (RAG) over traditional generation-only models. Lewis et al.[15] introduced the original RAG framework, building a hybrid architecture that combines a non-parametric document retriever with a generative model, showing that combining retrieval with generation improves both factual accuracy and citation of source documents compared to generation-only baselines such as BART. Building on this, Redelaar et al.[24] applied RAG to Dutch legal precondition questions, showing that retrieval of legislative fragments substantially improved citation accuracy as measured by the ALCE framework. Moreover, recently the LegalBench-RAG benchmark [22] was proposed to evaluate fine-grained retrieval in legal RAG pipelines. The benchmark emphasizes minimal snippet retrieval over full-document retrieval, addressing the risks of hallucination and context overflow. These studies show that RAG is particularly well-suited for high-stakes, citation-sensitive domains like law.

Hallucinations, where an LLM fabricates case names or citations are particularly problematic in legal NLP due to the need for strict factual accuracy. Huang et al.[10] provided a detailed classification of hallucination types, such as factual and faithfulness errors, and identified root causes which include context integrity and retrieval failures. The effectiveness of Retrieval-Augmented Generation (RAG) in reducing hallucinations depends on the quality and integration of retrieved content. Park et al.[21] propose LRAGE, a modular evaluation toolkit for legal RAG systems, which supports confidence-based filtering and post-generation citation verification. Their results further show that hallucinations often originate from low-relevance retrievals. These studies highlight that hallucination mitigation in legal RAG requires coordinated improvements in both retrieval and generation components.

The use of sequence-to-sequence models for legal document summarization has been explored by several recent studies. Schraagen et al.[25] used a BART model fine-tuned on Dutch Supreme Court rulings from `Rechtspraak.nl` to generate readable and grammatically correct summaries. However, human evaluation of outputs has revealed that the summaries often omitted key elements such as legal background, court considerations, or the final judgment, thus limiting the professional reliability. Bannihatti Kumar et al.[1] assessed the transferability of BART and T5 models across four legal domains and concluded that models trained on a single domain performed poorly when applied to out-of-domain legal texts. Even in multi-domain settings, generated summaries lacked structural grounding and citation precision. These studies highlight the core limitation: lack of factual anchoring which in turn reduces system trustworthiness.

Hallucinations of LLMs remain a persistent challenge in citation-sensitive domains like law: one strategy used to mitigate this is the use of a second-pass LLM reviewer who critiques and revises the initial output. While being an intuitively promising approach, the effectiveness of such reviewers is still contested, especially in vertical domain-specific applications. Evidence that second-pass refinement can outperform single-step generation in general-purpose tasks has been introduced by Madaan et al. [17] in *Self-Refine*, a framework in which the same LLM critiques and iteratively revises its own outputs without additional training or external supervision. It achieved an average absolute improvement of 20% in output quality across diverse tasks, with human annotators consistently preferring the refined versions. This suggests that LLMs, even in their original configuration, can benefit from structured self-evaluation loops.

In the framework *Chain-of-Verification* Dhuliawala et al. [7] explains that a four-step pipeline that prompts the model to generate verification questions about its own claims, answer them, and revise its initial response improves the relative F1 score by 23% for question answering. Highlighting that factual inconsistencies often go uncorrected in single-pass generation, but can be surfaced through self-interrogation.

Despite these advances, domain-specific evaluations reveal that hallucination risks remain high. In legal settings, Dahl et al. [3] found that state-of-the-art LLMs hallucinated legal facts or citations in 69–88% of cases, even when asked verifiable questions. Similarly, Magesh et al. [18] showed that retrieval-augmented legal research tools still produced "potentially insidious" hallucinations. These findings suggest that legal text generation may require more than just retrieval, it may require a robust, domain-aware second-pass verifier.

To summarize, prior research has explored legal-domain LLMs, factuality in RAG, and legal summarization, but no study has yet evaluated how a RAG-based system can be used specifically for generating structured legal memos in Dutch law, nor how hallucinations and citation quality can be systematically assessed in this setting. This gap highlights the novelty and relevance of the present study.

A.3 Dataset Construction

A.3.1 Post-Chunking Validation. Because of the importance of chunks in our RAG pipeline, a comprehensive validation of both locally generated and remotely stored chunks from `supabase` was executed. Each chunk was checked for essential metadata completeness, and its distribution of tokens to assess length uniformity and to detect anomalies such as empty or short chunks. Furthermore, the chronological representativeness was checked by extracting the official judgment data from each ruling and computing the corresponding calendar quarter, this was done to identify potential temporal imbalances in the dataset. A summary of key validation metrics such as chunk volume, average token length, section distribution, and quarterly case coverage can be found in Appendix A.8.

A.3.2 Embedding and Indexing. Using the `intfloat/multilingual-e5-large` model [30] each textual chunk was converted into a dense vector representation, the model was selected for its strong zero-shot performance across multiple languages and because produces high-quality passage-level embeddings. According to the E5 convention each chunk was prefixed with the word `passage:` before encoding. To maintain consistency and avoid memory bottlenecks during inference, embeddings were computed in mini-batches on the CPU and normalized to unit length post hoc.

The resulting embeddings, each associated with its corresponding chunk metadata, were stored in a Supabase-hosted PostgreSQL database enhanced with `pgvector` [14], a PostgreSQL extension for efficient vector similarity search. To optimize retrieval performance, the database used a multi-tiered indexing strategy, first an HNSW index with cosine distance metrics (m=16, ef_construction=64) accelerated approximate nearest neighbor searches over the 1024-dimensional embeddings, while the targeted GIN indices on JSONB metadata fields (ECLI, court, and legal section identifiers) enabled efficient structured filtering.

Inserts were performed in batches of 500 rows per API call to balance Supabase's throughput limits with transaction reliability. Each row contained the chunk's plaintext content, metadata (serialized as `jsonb`), and its vector representation, yielding a final corpus of 1,944 indexed entries. The hybrid indexing architecture allowed simultaneous semantic search via vector proximity and exact metadata filtering.

A.4 Legal Intake Form

Table 1. Structured Legal Intake Form for Memo Generation

Field	Description
Betwist besluit (Contested Decision)	What decision is being challenged (e.g., rejection of a benefits application).
Gewenst resultaat (Desired Outcome)	What the user hopes to achieve through the appeal or objection.
Kritieke feiten (Critical Facts)	Case-specific factual background that may influence the legal outcome.
Toepasselijke wet (Applicable Law)	The statute or regulation considered relevant to the dispute (e.g., AOW, WIA).
Doelgroep (Target Audience)	Whether the memo is intended for the client, a colleague, or another legal actor.

A.5 Example Structured Input

Table 2. Structured user input for legal memo generation

Veld (Field)	Voorbeeld (Dutch)	Translation (English)
Betwist besluit *(Contested Decision)*	Afwijzing AOW-aanvraag wegens onvoldoende opbouwjaren	Rejection of AOW pension application due to insufficient qualifying years
Gewenst resultaat *(Desired Outcome)*	Toekenning volledige AOW	Grant of full AOW pension
Kritieke feiten *(Critical Facts)*	SVB heeft buitenlandperiodes niet meegeteld, terwijl cliënt in die periode wel in Nederland verzekerd was via detachering en premies heeft betaald.	SVB did not count foreign periods, even though the client was insured in the Netherlands during that time via secondment and had paid contributions.
Toepasselijke wet *(Applicable Law)*	AOW	AOW (General Old Age Pensions Act)
Doelgroep *(Target Audience)*	Cliënt	Client

A.6 Example Retrieved Chunk Structure

```
{
  "id": "...",
  "chunk_index": ...,
  "sub_chunk_index": ...,
  "ecli": "...",
  "text": "...",
  "similarity": ...,
  "metadata": {
    "date": "...",
    "ecli": "...",
    "court": "...",
    "title": "...",
    "quarter": ...,
    "section": "...",
    "subject": "...",
    "procedure": "...",
  }
}
```

Listing 1. Example Retrieved Chunk Structure

A.7 Example Evaluation Object Structure


```
{
    "citation_precision": 1.0,
    "citation_recall": 1.0,
    "predicted_eclis": [
        "ECLI:NL:RBROT:2023:5868",
        "ECLI:NL:RBDHA:2020:2941",
        "ECLI:NL:RBZWB:2024:9",
        "ECLI:NL:RBOVE:2021:3751",
        "ECLI:NL:RBZWB:2023:4665"
    ],
    "reference_eclis": [
        "ECLI:NL:RBROT:2023:5868",
        "ECLI:NL:RBDHA:2020:2941",
        "ECLI:NL:RBZWB:2024:9",
        "ECLI:NL:RBOVE:2021:3751",
        "ECLI:NL:RBZWB:2023:4665"
    ],
    "fabricated_eclis": 0,
    "ungrounded_statements": 0,
    "ungrounded_sentences": [],
    "hallucinated": false,
    "threshold": 0.7,
    "similarity_metric": "cosine",
    "num_sentences": 29,
    "num_chunks": 6,
    "ungrounded_ratio": 0.0
}
```


Listing 2. Example Evaluation Object Structure

A.8 Chunk Dataset Validation Summary

Table 3. Chunk Dataset Validation Summary

Metric	**Local**	**Supabase**
Total chunks	1,944	1,944
Avg tokens/chunk	138	138
Token length violations	0	0
Unique legal cases (by ECLI):		
Total cases	200	
Cases/quarter (2020–2024)	10	
Chunk distribution:		
`OVERWEGINGEN` (Considerations)	1,569 (80.7%)	1,569 (80.7%)
`BESLISSING` (Decision)	247 (12.7%)	247 (12.7%)
`ABSTRACT`	128 (6.6%)	128 (6.6%)

A.9 Evaluation Summary (Without Reviewer Component)

Table 4. Evaluation results across thresholds and similarity metrics without the reviewer component. Metrics shown: citation precision, recall, F1 score, ungrounded ratio, and hallucination rate.

Metric	Thresh	Prec.	Recall	F1	Ungr. Ratio	Hall. Rate
cosine	0.6	1.00	0.78	0.87	0.00	0.00
	0.7	1.00	0.78	0.87	0.00	0.00
	0.8	1.00	0.78	0.87	0.08	1.00
	0.9	1.00	0.78	0.87	0.88	1.00
dot	0.6	1.00	0.78	0.87	0.00	0.00
	0.7	1.00	0.78	0.87	0.00	0.00
	0.8	1.00	0.78	0.87	0.08	1.00
	0.9	1.00	0.78	0.87	0.88	1.00
euclid.	0.6	1.00	0.78	0.87	0.02	0.40
	0.7	1.00	0.78	0.87	0.94	1.00
	0.8	1.00	0.78	0.87	1.00	1.00
	0.9	1.00	0.78	0.87	1.00	1.00

A.10 Evaluation Summary (With Reviewer Component)

Table 5. Evaluation results across thresholds and similarity metrics from the RAG pipeline with the reviewer component. Metrics shown: citation precision, recall, F1 score, ungrounded ratio, and hallucination rate.

Metric	Thresh	Prec.	Recall	F1	Ungr. Ratio	Hall. Rate
cosine	0.6	1.00	0.78	0.87	0.00	0.00
	0.7	1.00	0.78	0.87	0.00	0.00
	0.8	1.00	0.78	0.87	0.11	1.00
	0.9	1.00	0.78	0.87	0.77	1.00
dot	0.6	1.00	0.78	0.87	0.00	0.00
	0.7	1.00	0.78	0.87	0.00	0.00
	0.8	1.00	0.78	0.87	0.11	1.00
	0.9	1.00	0.78	0.87	0.77	1.00
euclid.	0.6	1.00	0.78	0.87	0.02	0.33
	0.7	1.00	0.78	0.87	0.81	1.00
	0.8	1.00	0.78	0.87	0.99	1.00
	0.9	1.00	0.78	0.87	1.00	1.00

A.11 F1 scores across temperatures (GPT-4.1 reviewer)

Table 6. F1 scores across temperatures (GPT-4.1 reviewer)

Metric	Temp Threshold	0.02	0.05	0.2	0.5	0.7	0.9
cosine	0.6	0.87	0.87	0.87	0.88	0.89	0.89
	0.7	0.87	0.87	0.87	0.88	0.89	0.89
	0.8	0.87	0.87	0.87	0.88	0.89	0.89
	0.9	0.87	0.87	0.87	0.88	0.89	0.89
dot	0.6	0.87	0.87	0.87	0.88	0.89	0.89
	0.7	0.87	0.87	0.87	0.88	0.89	0.89
	0.8	0.87	0.87	0.87	0.88	0.89	0.89
	0.9	0.87	0.87	0.87	0.88	0.89	0.89
euclidean	0.6	0.87	0.87	0.87	0.88	0.89	0.89
	0.7	0.87	0.87	0.87	0.88	0.89	0.89
	0.8	0.87	0.87	0.87	0.88	0.89	0.89
	0.9	0.87	0.87	0.87	0.88	0.89	0.89

A.12 F1 scores across temperatures (Claude Sonnet 4 reviewer)

Table 7. F1 scores across temperatures (Claude Sonnet 4 reviewer)

Metric	Temp Threshold	0.02	0.05	0.2	0.5	0.7	0.9
cosine	0.6	0.85	0.87	0.87	0.88	0.87	0.87
	0.7	0.85	0.87	0.87	0.88	0.87	0.87
	0.8	0.85	0.87	0.87	0.88	0.87	0.87
	0.9	0.85	0.87	0.87	0.88	0.87	0.87
dot	0.6	0.85	0.87	0.87	0.88	0.87	0.87
	0.7	0.85	0.87	0.87	0.88	0.87	0.87
	0.8	0.85	0.87	0.87	0.88	0.87	0.87
	0.9	0.85	0.87	0.87	0.88	0.87	0.87
euclidean	0.6	0.85	0.87	0.87	0.88	0.87	0.87
	0.7	0.85	0.87	0.87	0.88	0.87	0.87
	0.8	0.85	0.87	0.87	0.88	0.87	0.87
	0.9	0.85	0.87	0.87	0.88	0.87	0.87

B APPENDIX B

B.1 User Interface Promoting Explainable AI

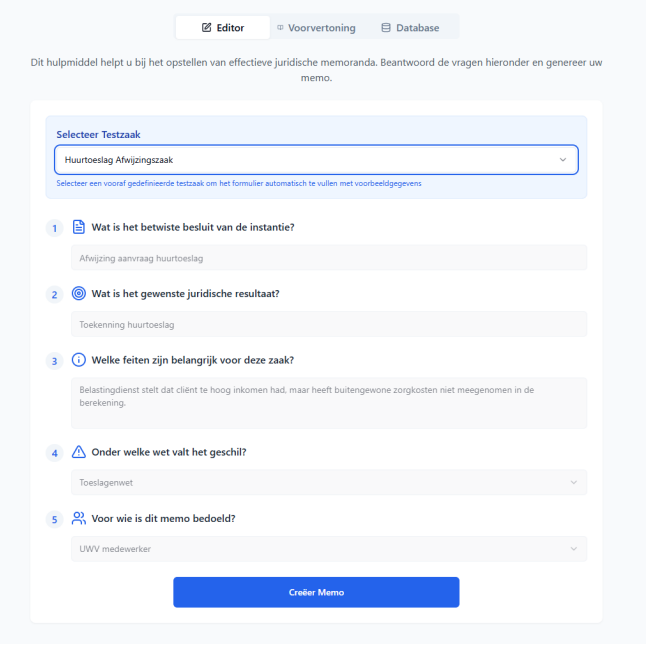


Fig. 7. Structured legal intake form used to collect key elements of the dispute, including the contested decision, desired outcome, critical facts, applicable law, and intended audience.

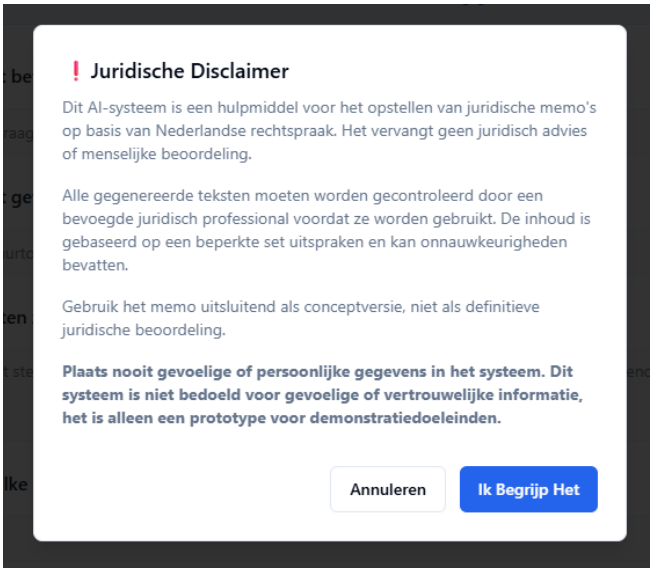


Fig. 8. Disclaimer step that informs users about the limitations of AI-generated legal memos and reinforces the human-in-the-loop principle before generation begins.

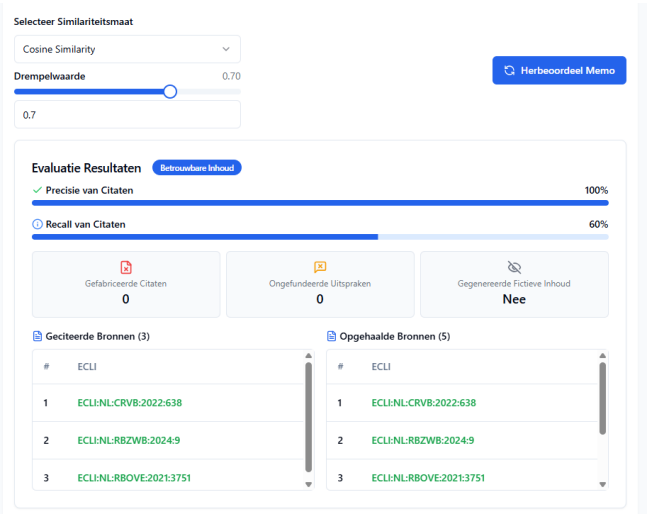


Fig. 9. Evaluation interface showing automated citation and hallucination metrics computed after memo generation. Used internally to track precision, recall, and semantic grounding.

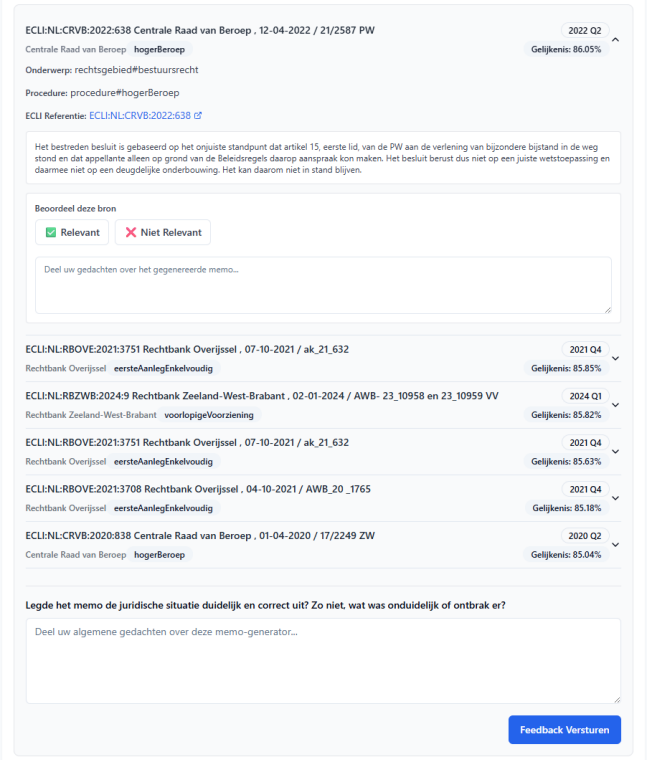


Fig. 10. Generated memo displayed alongside retrieved case law fragments (chunks), including ECLI identifiers and similarity scores, enabling traceability and citation verification.

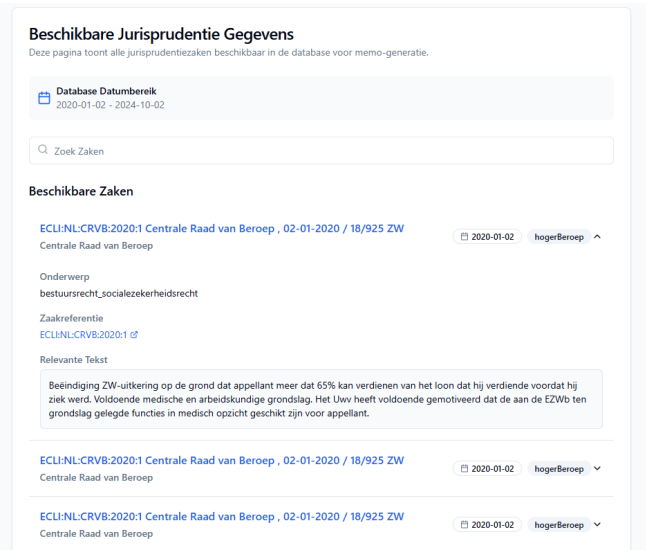


Fig. 11. Searchable database of jurisprudence used by the RAG system, allowing users to inspect which legal cases the model can retrieve during memo generation.

C APPENDIX C

C.1 Expert Feedback on Generated Memos

Table 8. Feedback from two legal professionals on the clarity, relevance, and legal grounding of AI-generated memos.

Reviewer	**Case Type**	**General Feedback: Legde het memo de juridische situatie duidelijk en correct uit? Zo niet, wat was onduidelijk of ontbrak er?**	**General Feedback: Did the memo explain the legal situation clearly and correctly? If not, what was unclear or missing?**
Legal Professional 1	WW	Ja op zich wel. Gaat erom dat het UWV ook de beoordeling van een onafhankelijke arts moet betrekken bij haar besluitvorming, als zo'n rapport er is en deze afwijkt van het oordeel van de verzekeringsarts. Het komt met een aantal relevante uitspraken van de CRvB waaruit dat blijkt. Of er ook uitspraken zijn waarin anders wordt geoordeeld wordt niet gegeven.	Yes, in principle. The point is that the UWV must also take into account the assessment of an independent doctor in its decision-making if such a report exists and it differs from the opinion of the insurance doctor. This is evident from a number of relevant rulings by the CRvB. Whether there are also rulings in which a different opinion is given is not stated.
Legal Professional 2	WIA	Ik vind de lay-out prettig en overzichtelijk. Het komt overeen met memo's/notities die ik schrijf (casusbeschrijving, wetgeving/jurisprudentie en toepassing, conclusie). De notitie is goed leesbaar en de casusbeschrijving was helder. De samenvatting van de jurisprudentie kan wat concreter. Die betreft nu niet altijd de kern van de zaak. Dat is ook lastig, want meestal behoeft dat duiding van bepaalde juridische zinnen/constructies. Maar wat ik heel mooi vind, is dat verschillende uitspraken bij elkaar worden genomen als ze enige overeenkomsten hebben. Wat verder opvalt is dat er veel uitspraken over de WIA in staan, terwijl de casus alleen over de WW gaat. Waar het op aanslaat waarschijnlijk is dat de werknemer tijdelijk ziek is geweest en dan snap ik wel dat het systeem richting arbeidsongeschiktheid gaat. De conclusie die er uit komt, is dan op zich wel weer juist voor zover ik begrijp en ook de aanbeveling is niet verkeerd. Ik ben wel benieuwd welke stappen het systeem heeft gemaakt om tot de conclusie te komen.	I find the layout pleasant and clear. It corresponds with the memos/notes I write (case description, legislation/case law and application, conclusion). The note is easy to read and the case description was clear. The summary of the case law could be a little more specific. It does not always concern the core of the case. That is also difficult, because it usually requires interpretation of certain legal phrases/constructions. But what I really like is that different rulings are grouped together if they have any similarities. What is also striking is that there are many rulings about the WIA, while the case only concerns the WW. What is probably relevant is that the employee was temporarily ill, and I understand that the system then moves towards incapacity for work. The conclusion that emerges is, in itself, correct as far as I understand, and the recommendation is not wrong either. I am curious to know what steps the system took to reach this conclusion.

Table 9. Full chunk-level feedback from Legal Professional 2 in Dutch with English translations. Each comment addresses the relevance and applicability of retrieved case law.

ECLI	Relevant	Reviewer Comment (Dutch)	Translation (English)
ECLI:NL:CRVB:2023:10	No	Hier betreft het een situatie waarin de vraag voorligt of er sprake is van nieuwe feiten en omstandigheden die een eerder genomen besluit - waartegen geen beroep is ingesteld - kunnen vernietigen. Dit is een meer juridisch-technische beoordeling dan een toetsing aan de wet. Ik zie daarom onvoldoende gelijkenissen met de onderhavige zaak.	This concerns a situation where the question is whether new facts and circumstances can overturn a previously made decision, one not appealed. This is more a technical legal assessment than a statutory test. I therefore see insufficient similarity with the present case.
ECLI:NL:CRVB:2021:761	No	De situatie van werk naar WW is een andere situatie dan van WW naar WIA. Aan de WIA zijn bepaalde eisen gesteld om aanspraak te kunnen maken op een uitkering op basis van die wet. Die eisen zijn anders dan van de WW. Het gaat hier om een situatie waarbij de werknemer tijdens zijn werkzame leven ziek is geweest en het UWV wil die weken niet meetellen als gewerkte weken. Dit is een andere situatie. Analoge toepassing is daarom lastig.	Transitioning from work to WW is different than going from WW to WIA. The WIA has different eligibility requirements than WW. In this case, the employee was sick while working, and the UWV doesn't want to count those weeks. This is a different scenario. Analogous application is therefore difficult.
ECLI:NL:CRVB:2021:768	No	Om dezelfde reden als de eerste en de tweede uitspraak.	For the same reasons as the first and second rulings.
ECLI:NL:CRVB:2020:1364	No	Dit gaat over wat je moet aandragen in beroep (grieven tegen de uitspraak van de rechtbank). Dat heeft de appellant niet gedaan waardoor een beroep niet kan slagen. Er heeft slechts een herhaling van argumenten bij de rechtbank plaatsgevonden. Ook hier is sprake van een juridisch-technische vraag (namelijk verschil tussen eerste aanleg en beroep). Ik zie daarom niet voldoende gelijkenissen met de onderhavige zaak.	This concerns what must be raised on appeal (objections against the court ruling). The appellant failed to do so, making the appeal unlikely to succeed. Only repeated arguments were made. This again is a technical issue (difference between first instance and appeal). I see insufficient similarity to the current case.
ECLI:NL:CRVB:2020:2340	No	Ik heb deze gekwalificeerd als niet relevant, maar ik snap bij deze wel waarom de uitspraak er tussen staat, omdat de situatie beter te vergelijken is met die van de onderhavige zaak. Ook hier geldt dat er andere voorwaarden zijn voor de WIA dan voor de WW, maar het idee van bewijsvoering komt hier nadrukkelijk naar voren en dat is wel belangrijk ook voor de onderhavige zaak.	I've marked this as not relevant, but I understand why it was included, it's more comparable to the present case. Although the requirements for WIA differ from WW, the issue of burden of proof is central, and that is relevant here.
ECLI:NL:CRVB:2020:2343	No	Hier geldt eigenlijk hetzelfde als bij de vorige uitspraak.	Same comment as the previous ruling.